

日 本 国 特 許 庁

PATENT OFFICE
JAPANESE GOVERNMENT



別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。

This is to certify that the annexed is a true copy of the following application as filed with this Office.

出 願 年 月 日
Date of Application:

1999年 8月27日

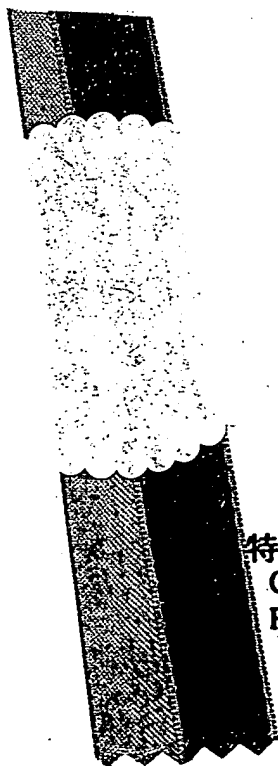
出 願 番 号
Application Number:

平成11年特許願第241057号

出 願 人
Applicant(s):

インターナショナル・ビジネス・マシーンズ・コーポレイション

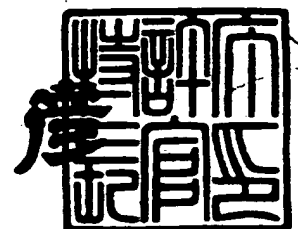
CERTIFIED COPY OF
PRIORITY DOCUMENT



1999年11月26日

特許庁長官
Commissioner,
Patent Office

近 藤 隆 彦



【書類名】 特許願

【整理番号】 JA999118X

【あて先】 特許庁長官 殿

【国際特許分類】 G06F 15/00

【発明者】

【住所又は居所】 神奈川県大和市下鶴間 1 6 2 3 番地 1 4 日本アイ・ピー・エム株式会社 東京基礎研究所内

【氏名】 那須川 哲哉

【発明者】

【住所又は居所】 神奈川県大和市下鶴間 1 6 2 3 番地 1 4 日本アイ・ピー・エム株式会社 東京基礎研究所内

【氏名】 長野 徹

【特許出願人】

【識別番号】 390009531

【住所又は居所】 アメリカ合衆国 1 0 5 0 4、ニューヨーク州アーモンク
(番地なし)

【氏名又は名称】 インターナショナル・ビジネス・マシーンズ・コーポレーション

【代理人】

【識別番号】 100086243

【弁理士】

【氏名又は名称】 坂口 博

【連絡先】 0 4 6 2 - 7 3 - 3 3 1 8、3 3 2 5、3 4 5 5

【選任した代理人】

【識別番号】 100091568

【弁理士】

【氏名又は名称】 市位 嘉宏

【先の出願に基づく優先権主張】

【出願番号】 平成11年特許願第192618号

【出願日】 平成11年 7月 7日

【手数料の表示】

【予納台帳番号】 024154

【納付金額】 21,000円

【提出物件の目録】

【物件名】 明細書 1

【物件名】 図面 1

【物件名】 要約書 1

【包括委任状番号】 9304391

【包括委任状番号】 9304392

【プルーフの要否】 要

【書類名】 明細書

【発明の名称】 データ分析システム

【特許請求の範囲】

【請求項 1】

データから特徴的な概念を取り出す、データ分析システムであって、

(1) 文書データを含むデータからカテゴリ別の概念を抽出する、概念抽出手段と、

(2) 前記カテゴリ別の概念において、同一カテゴリに属する概念のうち、対応する別のカテゴリに属する概念の中で占める割合が既定値を超えている概念を抽出する、特徴的概念抽出手段、

を有する、データ分析システム。

【請求項 2】

前記概念抽出手段 (1) が、

データ中の前記文書データを形態素解析する手段と、

前記形態素解析の結果に基づき、前記文書データの文節を生成する手段と、

前記文節内のキーワードを概念として抽出する手段であって、前記文節に対してカテゴリ辞書を適用して、文節内のキーワードに対して概念 (該キーワードの代表的意味を表現した置き換え表現) とカテゴリを付加する手段と、

前記文節を有する文を構文木生成規則に従い構文を解析する、構文解析手段と、

前記文節内の概念およびカテゴリを付加されたキーワードに関し、同一文中でのキーワードの係り受け関係を抽出する、係り受け抽出手段と、

前記カテゴリ別の概念を抽出する手段であって、前記キーワード間の係り受け関係に基づき、係り受けの関係にある各概念のカテゴリの組み合わせを抽出する手段と、

を含む、請求項 1 記載のシステム。

【請求項 3】

前記特徴的概念抽出手段 (2) が、

ユーザの命令を受け取る入力手段と、前記ユーザの命令を解析する、命令解析手段と、前記解析された命令に従い、前記カテゴリ別の概念を提示し、同一カテゴ

りに属する概念のうち、対応する別のカテゴリに属する概念の中で占める割合が既定値を超えている概念を他の概念と異なる属性で表示する手段とを有する、請求項 1 記載のシステム。

【請求項 4】

前記特徴的な概念を検出する手段（2）が、さらに、
抽出された概念の相対頻度を計算する手段と、
抽出された概念の集合から概念を検索する手段と、
カテゴリ別概念の頻度を計算する手段と、
得られた前記相対頻度、検索結果、概念の頻度を表示する手段と、
を含む、請求項 3 記載のシステム。

【請求項 5】

データから特徴的な概念を取り出す、概念抽出方法であって、
データ中の文書データに対して形態素解析を行う段階と、
前記形態素解析の結果に基づき、前記文書データの文節を生成する段階と、
前記文節に対してカテゴリ辞書を適用して、文節内のキーワードに対して概念（
該キーワードの代表的意味を表現した置き換え表現）とカテゴリを付加する段階
と、
前記文節を有する文を構文木生成規則に従い構文木を生成する段階と、
前記文節内のカテゴリを付加されたキーワードに関し、同一文中でのキーワード
の係り受け関係を抽出する段階と、
前記キーワード間の係り受け関係に基づき、係り受けの関係にある各概念のカテ
ゴリの組合せを抽出する段階と、
を有する、概念抽出方法。

【請求項 6】

データから特徴的な概念を取り出すプログラムを記録したコンピュータ読み取り
可能な記録媒体であって、該プログラムが、コンピュータに、
データ中の文書データに対して形態素解析を行う機能と、
前記形態素解析の結果に基づき、前記文書データの文節を生成する機能と、
前記文節に対してカテゴリ辞書を適用して、文節内のキーワードに対して概念（

該キーワードの代表的意味を表現した置き換え表現) カテゴリを付加する機能と、
 前記文節を有する文を構文木生成規則に従い構文木を生成する機能と、
 前記文節内のカテゴリを付加されたキーワードに関し、同一文中でのキーワードの係り受け関係を抽出する機能と、
 前記キーワード間の係り受け関係に基づき、係り受けの関係にある各概念のカテゴリの組合せを抽出する機能と、
 を実現させる記録媒体。

【発明の詳細な説明】

【0001】

【産業上の利用分野】

本発明は、大量のデータから特徴的な概念を抽出する技術に関し、特に同一カテゴリを付加したの概念の比較により、大量のデータから注目に値する有効な知識を獲得する方法およびシステムに関する発明である。

【0002】

【従来の技術】

ある製品に関して製造元の会社によせられる顧客からの苦情、問題、意見などの情報は従来何らかの形で文書データとして記憶されている。この文書データは異なる顧客から寄せられるのでその内容も異なる。この文書データが少ない場合には人手により解析することは容易であった。現在は製品の電話サポートなどのように多くの不特定の顧客からのコメントが容易に電子化されて記憶されるので集められる文書データは莫大なものとなり人手で解析できる範囲をはるかに超えている。そこで大量の文書の内容を様々な形で視覚化し分析に役立てようという試みが次第に増えてきている。しかし、従来の方法では、名詞句を中心としたキーワードを抽出し、キーワードの分布を表示（例えば、キーワード間の相関関係を計算し、その結果に基づいて相関の強いキーワードをクラスタ化して表示）する程度の処理しか行っていない。その程度の処理の結果を用いた場合には、ユーザーが様々な観点からデータを絞り込みながら、主観的に注目に値すると思える部分を発見するまで手探りで分析するしかなく、分析過程におけるユーザーの負

担が大きい。また、多種多様なキーワードを混合して、まとまりの無いままの状態で扱うため、有効な結果にたどり着くのは難しい。

つまり、テキストをデータマイニングの手法（クラスタリングや相関ルールの分析など）で解析してみようという試みは、データマイニングが注目され始めた頃から存在するものの、従来の手法では、テキストから抽出した分析単位が、単なる文字列としてのキーワードに過ぎなかった為、有効な結果を得られずに終わってしまうケースが多い。

【0003】

【発明が解決しようとする課題】

従って、本発明が解決しようとする課題は、大量データから特徴的な概念を獲得する方法およびシステムを提供することである。

また別の課題は、大量データから注目に値する概念を自動的に見つけ出す方法およびシステムを提供することである。

また別の課題は、大量のデータから有効な特徴的な概念を獲得するにあたり、ユーザビリティに優れた分析方法およびシステムを提供することである。

【0004】

【課題を解決するための手段】

上記課題を解決するために、本発明は文書を含む大量のデータから特異な特徴を有する概念を抽出することにより、有効な知識を獲得する方法およびシステムを提供する。

本発明は、概念抽出装置と特徴的概念抽出装置とを有する。前記概念抽出は文書データを含むデータからカテゴリ別の概念を抽出する。前記特徴的概念抽出装置は前記抽出した概念の中から特徴的な概念を抽出する装置であって、前記カテゴリ別の概念において、同一カテゴリに属する概念のうち、対応する別のカテゴリに属する概念の中で占める割合が既定値を超えている概念を抽出する。

【0005】

前記概念抽出装置では、語彙辞書や文法知識を利用した形態素解析やカテゴリ辞書を用いた曖昧性解消手法を用いて、非定型テキストから、カテゴリ別の概念を抽出する。前記特徴的概念抽出装置では、カテゴリとカテゴリの各組み合わせ

において、同じカテゴリの概念のうち、対応する別のカテゴリの概念の占める割合の比や差が既定値（しきい値）を超えているものを見つけ出す。さらにカテゴリとカテゴリの各組み合わせを表形式で表示し、該表中において注目すべき項目を他と異なる属性で表示したり、リストアップして提示するなどの処理を行う。

【請求項 7】

より詳細には、前記概念抽出装置は、データ中の前記文書データを形態素解析する手段と、前記形態素解析の結果に基づき、前記文書データの文節を生成する手段と、前記文節内のキーワードを概念として抽出する手段であって、前記文節に対してカテゴリ辞書を適用して、文節内のキーワードに対して概念（該キーワードの代表的意味を表現した置き換え表現）とカテゴリを付加する手段と、前記文節を有する文を構文木生成規則に従い構文を解析する、構文解析手段と、前記文節内の概念およびカテゴリを付加されたキーワードに関し、同一文中でのキーワードの係り受け関係を抽出する、係り受け抽出手段と、前記カテゴリ別の概念を抽出する手段であって、前記キーワード間の係り受け関係に基づき、係り受けの関係にある各概念のカテゴリの組み合わせを抽出する手段から構成される。

前記特徴的概念抽出装置は、ユーザの命令を受け取る入力手段と、前記ユーザの命令を解析する、命令解析手段と、前記解析された命令に従い、前記カテゴリ別の概念を提示し、同一カテゴリに属する概念のうち、対応する別のカテゴリに属する概念の中で占める割合が既定値を超えている概念を他の概念と異なる属性で表示する手段から構成される。

【0006】

【発明の実施の形態】

図 1 は、本発明のデータ分析システム概要を示すブロック図である。なお実施例としてコンピュータ製品に関して電話により問い合わせされたデータを例にして、これを解析し、特徴ある概念を抽出するデータ分析システムについて説明する。このデータ分析システムは顧客から寄せられた大量の問い合わせデータ 150 を機械的に解析できるように予めデータ変換部 110 で解析できる状態に変換する。この変換されたデータからカテゴリ辞書 170 を用いて概念付きデータを抽出し、抽出された概念付きデータから特徴的な概念を検索・検出する手段を提供

する。

【0007】

より詳細には、図1のブロック110は問い合わせデータ150を入力としてラベル付きデータ160を出力するデータ変換部である。問合せデータ150に含まれる非定型のデータと、同じく含まれる定型のデータを同一に保持する形式のデータ（ラベル付きデータ）を作成する。ここで定型データとは、主に複数の項目情報などからなり、各項目情報がどの位置から開始され、何文字まで許されるかなど予め決められており、解析の容易なデータ形式である。逆に非定型データは主に文章などのようにその長さがまちまちな情報であり解析の難しい形式である。またブロック120は、入力されたラベル付きデータのキーワードに対してカテゴリ辞書170を用いてカテゴリを付加し、カテゴリが付加されたキーワードのうち、同じ文中で係り受け関係のあるものを、より具体的な意味を表現する概念（ラベル・概念付きデータ180）として抽出する、概念抽出部である。

【0008】

ここで概念とは「キーワード」に「カテゴリ」を付加したものの意味し、さらに該概念の係り受けによる組み合わせを複合概念（さらに上位の「概念」）として抽出する。また「ラベル」は「カテゴリ」とデータの属性を含む。

【0009】

ブロック130は、ラベル・概念付きデータ180を入力として、特徴のある概念を検索・抽出する検索・特徴検出部130である。このブロックは検索・特徴抽出を効率良く行えるようにするため、データ全体もしくは部分集合における頻度分布などの統計情報を作成・保持する。同じカテゴリに属する概念は同じような振る舞いをする（同じような出現傾向を持ち、同じような概念と共起する）傾向があると想定し、各概念の振る舞いを全体、もしくは部分集合の値と比較することで、効率良く、注目すべき情報を検索・検出する。さらにこの検索・特徴検出部130は、統計情報を視覚的に表示する機能を有し、特徴ある概念の分布差異の提示を行う。

【0010】

上記各ブロック110、120、130について以下に詳細に説明する。

〔データ変換部 1 1 0〕

まずブロック 1 1 0 に入力される問い合わせデータの例は以下のようなものである。

【 0 0 1 1 】

----- 問い合わせデータ例 -----

1999/01/01

0000001

タイトル：ノートで日本語が使えない

マシンタイプ：製品 A

問題種別：総合案内

C A L L 種別：案内

回答・対応種別：窓口対応

解決期間：1 日

通話時間：2 1 分

ご質問：ノートパソコンで日本語が使えないので、OS を再インストールしたのですが、それ以降、MODEM とイーサネットカードが使えなくなっていました。

----- 問い合わせデータ例 -- 終わり -----

【 0 0 1 2 】

このように、問合せデータは日付や問い合わせ番号、項目名、項目の内容（予め既定された選択肢からの選択、連続値や離散値、自由に記入できるタイトルなど）、文章からなる質問内容など、さまざまな形式の項目から構成された定型データ、非定型データの混在からなる。

【 0 0 1 3 】

ブロック 1 1 0 は上記のような問い合わせデータ 1 5 0 を入力とし、下記のようなラベル付きデータに変換する。

【 0 0 1 4 】

----- ラベル付きデータ -----

ID199901010000001

TIノートで日本語が使えない

KWM1MT : 製品A

KWQ3TC : 総合案内

KWQ4TD : 案内

KWQ2PT : 窓口対応

KWP3SD : 1日

KWP4CM : 21分

CTQ : ノートパソコンで日本語が使えないので、OSを再インストールしたのですが、それ以降、MODEM とイーサネットカードが使えなくなっていました。

----- ラベル付きデータ --終わり----

【0015】

このように、非定形データを含む問い合わせデータを上記のようなラベル付きデータに変換することにより、さまざまな型のデータを同じ形式に変換する。上の例ではTIはタイトル、CTは元の問合せ内容を示し、KW+2byteは項目の種類を示す。この2バイトがカテゴリを表わし、Q3は問題種別、Q4はCALL種別、Q2が回答・対応種別、P3は解決期間、P4は通話時間を示す。ここでラベルとは項目内容のカテゴリを示す情報で”KWM1MT :”のようにカテゴリを含む長さ固定の情報である。図2にデータ変換部のフローチャートを示す。ステップ210で問い合わせデータ150を読み込み、ステップ220でデータの終了であるかを判断し、そうでなければステップ230で形式を変換する。データが終了した場合はステップ240で変換を終了する。例えば問い合わせデータの”タイトル :”を発見した場合にはそれを”TI”に変換し、その横に項目内容”ノートで日本語が使えない”を記録するという具合である。このようにデータ変換部では、問い合わせデータを概念抽出の前処理としてデータを解析しやすいようにラベルを付加して形を整える。当業者であればこのような変換は変換規則を変更することによりあらゆるデータに対して変換可能であることが容易に推察されるであろう。

【0016】

〔概念抽出部120〕

次にブロック330は、ブロック320で生成された文節に対してカテゴリ辞書

340を適用し文節内のキーワードに対してカテゴリを付加する辞書適用装置である。キーワードは文節内の一纏りの文字列である。上記カテゴリ辞書340の構造の例を以下に示す。

【0017】

----- カテゴリ辞書の例 -----

元表現	品詞	概念（置き換え表現）	カテゴリ
ノートパソコン	固有名詞	ノートパソコン	N1
OS	固有名詞	オペレーティングシステム	N2
壊れる	動詞	故障する	VC

----- カテゴリ辞書の例 - 終わり ---

【0018】

カテゴリ辞書340は〔元表現 品詞 概念 カテゴリ〕の組み合わせから構成されている。ここで元表現は文書データ中のキーワードに相当し、品詞はそのキーワードの種別、概念はそのキーワードの置き換え表現であり、上記でいえばキーワード”OS”は該キーワードの代表的意味を表現した置き換え表現として”オペレーティングシステム”として統一される。最後にカテゴリはそのキーワードの性質を有するより大きなグループを意味する。上記辞書においてカテゴリは、N1=ハードウェアを意味する、N2=ソフトウェアを意味する、VC=問題を意味する、というように対応付けられている。キーワードにカテゴリを付加することで単なる文字列ではなく意味を持った概念として扱うことができる。（例えば、「ワシントン」という文字列を単なるキーワードとして扱えば人名か地名か区別できず有効な分析が行えないが、〔人名〕〔地名〕といったカテゴリを付加することで、意味を持つことが出来る）。体言（名詞類の語）に関しては上記カテゴリ辞書を参照してカテゴリを付加する。述語に関しては体言と同様にカテゴリ辞書を用いると共に、付属語の情報から〔問題〕〔要望〕〔疑問〕といったカテゴリ付けを行う。例えば、「壊れる」という動詞は（壊れる〔動詞〕故障する〔問題〕）というデータがカテゴリ辞書にあれば、〔問題〕というカテゴリに属する「故障する」という概念として抽出されるが、「…できない」「…したい」という表現は、製品

に関して電話により問い合わせされたデータである事を考えれば、それぞれ問題であること、要望であることが自明であるので、カテゴリ辞書を参照せずに [問題] [要望] として解釈することができる。

【 0 0 1 9 】

ブロック 3 5 0 はブロック 3 3 0 でキーワードに対してカテゴリが付加された文節を有する文を簡単な構文木生成規則で構文木を生成する構文木解析装置である。

【 0 0 2 0 】

ブロック 3 6 0 は文節内のカテゴリを付加されたキーワードのうち同じ文中で係り受け関係のあるものを、より具体的な意味を表現する概念として抽出する係り受け抽出装置である。このブロック 3 6 0 では構文解析装置 3 5 0 の構文解析結果により得られるキーワード間の係り受け関係に基づき、係り受けの関係にある各キーワードのカテゴリの組合せを概念（ラベル・概念付きデータ 3 7 0）として抽出する。ラベル・概念付きデータ 3 7 0 の例を以下に示す。

【 0 0 2 1 】

----- ラベル・概念付きデータの例 -----

ID199901010000001

TI ノートで日本語が使えない

KWM1MT : 製品 A

KWQ3TC : 総合案内

KWQ4TD : 案内

KWQ2PT : 窓口対応

KWP3SD : 1 日

KWP4CM : 2 1 分

CTQ : ノートパソコンで日本語が使えないので、OS を再インストールしたのですが、それ以降、MODEM とイーサネットカードが使えなくなっていました。

KWN1 ノートパソコン

KWN0 日本語

KWV2 使えない

KWW6ノートパソコン…使えない

KWN2OS

KWV6再インストールする

KWWDOS…再インストールする

KWN1MODEM

KWN1イーサネットカード

KWV2使えない

KWW6モデム…使えない

KWW6イーサネットカード…使えない

----- ラベル・概念付きデータの例 --終わり---

【0022】

上記のようにラベル・概念付きデータ370は、データ変換部で得られたラベル付きデータ160に、概念抽出部120で抽出されたデータを追加した形を取り、ラベル付きデータ160と同一形式のデータとなる。

【0023】

図12に本発明の概念抽出部の流れを実際の文に基づき説明する。まずステップ1210において、“MODEMとイーサネットカードが使えない。”という入力文が入力された場合、ステップ1220において、形態素解析装置310により、文が単語に区切られ、各単語に品詞番号が付加される。これにより“MODEMとイーサネットカードが使えない。”という入力文は、以下のように変換される。

[MODEM, 104] [と, 81] [イーサネット, 104] [カード, 104] [が, 75] [使, 10] [え, 44] [ない, 51] [。 , 100]

上の例では、104…固有名詞、81…格助詞「と」、75…格助詞「が」、10…動詞語幹、44…形容詞假定形活用語尾、51…打ち消しの助動詞「ない」、100…句読点、を表わしている。

【0024】

次にステップ1230、およびステップ1240で文節生成が行われる。ステップ1230の文節生成1では、形態素解析された文の単語列を文節にまとめる操作を行う。あらかじめ“{81, 75, 100, …}で文節を区切る”というルールを用

意しておき、このルールを文頭から適用して、文頭から文節ごとに区切っていく操作を行う。上記入力文の場合、先頭から3文節あるが、各文節の先頭の単語がそれぞれ、名詞・名詞・動詞であることから、それぞれ順に、体言句・体言句・用言句であると判断される。その結果、入力文は以下のように変換される。

【 0 0 2 5 】

{[MODEM, 104] [と, 81]}
 {[イーサネット, 104] [カード, 104] [が, 75]}
 {[使, 10] [え, 44] [ない, 51] [。 ,100]}

【 0 0 2 6 】

次にステップ 1 2 4 0 の文節生成 2 は、文節生成 1 で区切られた文節に対して、すべての文節を自立語と付属語の組にする。体言句は、名詞が複数含まれる場合は先頭から名詞を結合するようにする。例えば、{[イーサネット, 104] [カード, 104] → [イーサネットカード, 104]} のようにする。その後自立語の品詞コードを、一般名詞句を示す N1 に書き換える。用言句は、付属語列（[え, 44] [ない, 51] [。 ,100]）を解析し、否定の情報を示す [ない, 51] を取り出し、動詞の語幹 [使, 10] は [え, 44] と結合して、終止形「使える」にする。品詞コードを一般動詞を示す V1 にし、否定の情報を付加して -V1 とする。その結果、入力文は以下のように変換される。

【 0 0 2 7 】

{[MODEM, N1] [と, 81]}
 {[イーサネットカード, N1] [が, 75]}
 {[使える, -V1] [。 , 100]}

【 0 0 2 8 】

次にステップ 1 2 5 0 で、カテゴリ辞書を用いて、自立語と付属語の組に分解された文節に対してカテゴリ付を行う。ここで適用される辞書は、以下の3つである。

【 0 0 2 9 】

(MODEM N1 モデム NA)
 (イーサネットカード N1 イーサネットカード NA)

(使える -V1 使えない VC)

【0030】

なおNAはハードウェアを意味し、VCは問題を意味する。その結果、入力文は以下のように変換されてカテゴリが付された3つの文節が生成される。

【0031】

{[モデム, NA] [と, 81]}

{[イーサネットカード, NA] [が, 75]}

{[使えない, VC] [。 , 100]}

【0032】

次にステップ1260で、カテゴリ付された文節からなる文に基づき構文木を生成する。この時の係り受けルールの形式は（係り受け元文節の自立語，係り受け元文節の付属語，係り受け先文節の自立語，係り受け先文節の付属語）となっている。このルールを文頭の文節1 {[モデム, NA] [と, 81]} から適用する。一般にn番目の文節に対してn+1番目から最終N番目までの文節に係り受けのルールを適用する($n = 1 \sim N-1$)。係り受けルール中に(NA, 81, VC, *)というルールがあるので、{[モデム, NA] [と, 81]}と{[使えない, VC] [。 , 100]}に係り受けの関係があると判断される。なおルール中の*はすべての品詞またはカテゴリにマッチするという意味である。これを($n = 1 \sim N-1$)で行い、係り受け情報を含む文節を有向グラフとして表し、(係り受け元の文節番号，係り受け先の文節番号，自立語，カテゴリ，付属語の品詞番号)という形式に変換する。その結果、入力文は以下のように変換される。

【0033】

(1, 3, "モデム", NA, 81)

(2, 3, "イーサネットカード", NA, 75)

(3, NULL, "使えない", VC, 100)

なおNULLは係り受け先がないことを示す。

【0034】

最後に、ステップ1270において、構文解析された文を入力として、係り受け抽出ルールに従って係り受けを抽出する。抽出ルールは任意の長さのカテゴリの

列からなる。例えば(カテゴリ 1, カテゴリ 2, ..., カテゴリ n) という形式である。文節番号 1 から N までの係り受けを見て、係り受け抽出ルールに (NA, VC) というルールがあるので、「モデム…使えない」と「イーサネットカード…使えない」という 2 つの係り受けが抽出される。

結局、元の文書である「MODEM とイーサネットカードが使えない」という文章から

【 0 0 3 5 】

「モデム …使えない」 [ハードウェア…問題]

「イーサネットカード …使えない」 [ハードウェア…問題]

【 0 0 3 6 】

という概念情報が取り出されたことになる。このようにして取り出された概念情報は、ラベル・概念付きデータベース 1 8 0 に登録される。

【 0 0 3 7 】

図 4 に、本発明の概念抽出部 1 2 0 の処理のフローチャートを示す。

ステップ 4 2 0 でラベル付きデータ 1 6 0 内の文章 T を形態素 $W_0 \sim W_m$ に分割する。ここで形態素 W は、文字列 w と品詞 p で表される。すなわち $W = \{ w, p \}$ である。(以上が形態解析装置の処理である)

【 0 0 3 8 】

次にステップ 4 3 0 で全単語が文節に変換されたかどうかを判断し、そうであれば処理はステップ 4 4 0 へ移り、そうでなければステップ 4 3 2 で単語 W_n が付属語または句読点かどうかを判断し、その結果が No であればステップ 4 3 4 で、文節 P_i に単語 W_n を追加する。ここで文節 P は、1 つ以上の連続した単語 W の集合で $P = \{ W^* \} = \{ \{ w, p \}^* \}$ である。そして処理はステップ 4 3 0 に戻る。ステップ 4 3 2 の判断の結果が Yes であればステップ 4 3 6 で文節 P_i に単語 W_n を追加した後、その次の文節を用意する。処理はその後ステップ 4 4 0 に移る。ステップ 4 4 0 で全文節に対して処理が行われたかどうかを判断し、処理が終了していれば処理はステップ 4 5 0 に移り、そうでなければステップ 4 4 2 において、文節を $P = \{ \{ w, p \}^* \}$ の形式から $P' = \{ \{ w_1, p_1 \} \{ w_2, p_2 \} \}$ に変換する(ここで w_1 は自立語 w_2 は付属語である)。例えば $P = \{ [国際,$

名詞] [情勢, 名詞] [は, 助詞] } であれば、名詞句はひとつにまとめ、 $P' = \{ [国際情勢, 名詞] [は, 助詞] \}$ とする。そして処理はステップ 440 へ戻る。
 (以上が文節生成装置 320 の処理である)

【0039】

ステップ 450 では、自立語に対して辞書引きが終了したかどうかを判定する。その判定結果が No であれば、ステップ 452 で、 $((w1, p1) == (wa, pa))$ のエントリに対し $(w1 = wb, p1 = pb)$ とする。同時に用言の処理を行う。この時利用される辞書 454 は $[wa\ pa\ wb\ pb]$ のエントリの集合からなり、それぞれ [元表現, 品詞, 概念, カテゴリ] を表す。ここでの概念というエントリは、置き換え表現を意味する。元表現が「PC」であれば置き換え表現は「パソコン」となる。また全エントリはハッシュ構造で格納されているので高速にアクセスできる。例えばエントリ集合は、[マシン, 名詞, 機械, ハードウェア]、[壊れる, 動詞, 壊れる, 問題] 等である。この辞書を用いて、名詞句 $P' = \{ [マシン, 名詞] [が, 助詞] \}$ であれば、カテゴリ「問題」を逆の意味を表わすカテゴリ「好評」に変換し $P'' = \{ [機械, ハードウェア] [が, 助詞] \}$ となる。用言の句 $P' = \{ [壊れる, 動詞] [ない, 助詞] \}$ であれば、動詞→問題となるはずだが、「ない」があるので、 $P'' = \{ [壊れない, 好評] [NULL, NULL] \}$ とする。また辞書に該当するエントリがなければ何もしない。ステップ 450 の判定結果が Yes であれば処理はステップ 460 へ進む。(以上が辞書適用装置 330 の処理である)

【0040】

ステップ 460 で、構文木が完成したかどうか判定される。構文木が完成した場合、処理はステップ 470 へ進む。構文木が完成していない場合、処理はステップ 462 で、一般的に行われる構文解析を行い、結果として、 P_n と P_k がリンクされる。(以上が構文解析装置 350 の処理である。)

【0041】

ステップ 470 で、係り受けの抽出が終了したかどうかを判断する。もし終了していなければ処理はステップ 472 へ進み、ルールに基づき、 P_n に対してリンクされた 2 項関係を抽出し、ラベル・概念付きデータベース 180 に登録する。こ

の時、係り受けルール 474 を参照する。係り受けルール 474 のルールは [px py] のエントリの集合からなり、それぞれ [係り受け元カテゴリ, 係り受け先カテゴリ] を表す。例えば、 $P_n = \{ [\text{機械}, \text{ハードウェア}] [\text{が}, \text{助詞}] \}$, $P_k = \{ [\text{壊れる}, \text{問題}] [\text{NULL}, \text{NULL}] \}$ (nとkには係り受け関係があるものとする) であれば、上のルールを用いて、[ハードウェア, 問題] \rightarrow [機械, 壊れる] が抽出され、ラベル・概念付きデータベース 180 に登録される。ステップ 470 の判断が Yes であれば処理はステップ 480 で終了する。(以上が係り受け抽出装置 360 の処理である。)

【0042】

[検索・特徴抽出部 130]

図 5 に検索・特徴抽出部 130 のブロック図を示す。検索・特徴抽出部 130 は入力(命令)装置 570、表示部 510、概念検索装置 540、命令解析装置 520、カテゴリ別概念頻度計算装置 550、相対頻度計算装置 530 の各ブロックから構成される。なお概念検索装置 540 及びカテゴリ別概念頻度計算装置 550 は、ラベル・概念付きデータベース 560 にアクセスして概念情報の検索を行う。好適にはラベル・概念付きデータベース 560 においては、ラベル・概念付きデータに対してインデックスを生成し、高速な検索が可能な状態にしておく。

【0043】

命令解析装置 520 は、入力(命令)装置 570 から受け取る命令を解析して、各装置にパラメータとしての概念を送る。入力(命令)装置 570 は、図 14 におけるキーボード 6、マウス 7 などに相当し、ユーザの指示に従い、データ分析システムに対して所望の検索、表示を行わせるために使用される。相対頻度計算装置 530 は、文書の全体、または部分集合に対して相対頻度を計算する装置である。ここで相対頻度とは、全体または任意の集合 X に対して含まれる各概念と、任意の集合 Y に含まれる概念の集合を比較することにより計算される。

【0044】

概念検索装置 540 は、命令解析装置 520 の出力である概念または概念の組を入力として、文書の全体、または部分集合に含まれる概念の数と概念を含む文書

の I D を得る装置である。該装置により概念を含む文書集合を絞り込むことが出来る。

【 0 0 4 5 】

カテゴリ別概念頻度計算装置 5 5 0 は、命令解析装置 5 2 0 の出力に従い、文書全体、または部分集合に含まれる概念の数をカテゴリ別に、頻度順に得る装置である。以下に該装置の出力例を示す。（下の例で INPUT ではカテゴリを指定、N1 は [ハードウェア] を表すカテゴリであり、OUTPUT では [キーワード 出現頻度] となっている。

【 0 0 4 6 】

----- カテゴリ別キーワード頻度計算装置の出力例 -----

[INPUT] CATEGORY	N1
[OUTPUT] ハードディスク	2033
[OUTPUT] モニタ	1432
[OUTPUT] プリンタ	1001
[OUTPUT] モデム	420
[OUTPUT] スキャナ	212
[OUTPUT] イーサネットカード	143
[OUTPUT] マウス	3

----- カテゴリ別キーワード頻度計算装置の出力例 --終わり--

【 0 0 4 7 】

表示部 5 1 0 は図 6 に示す表示領域 (1) と図 7 に示す表示領域 (2) を含む G U I 画面から構成される。ユーザが表示部 5 1 0 中に表示された種々の項目を入力 (命令) 装置 5 7 0 により適宜選択したり、検索のためのパラメータなどを入力することにより、種々の結果 (頻度表示、検索結果表示など) を表示部 5 1 0 に表示する。例えば図 6 は、カテゴリ A を横軸、カテゴリ B を縦軸にとった 2 次元の表である。カテゴリ A の列に対して特徴的なカテゴリ B を表すセルが他のセルとは異なる属性 (高輝度表示、異色表示、点滅、網掛けなど) で表示されている (例えば概念 A3 の中で特徴的な概念は B1 である。また複数の概念が特徴的なこともある) 。上記異なる属性で表示されたセルをクリックすることで、概念 A x と

概念Byを含む集合の検索が行われ、そこで得られた集合に対しても再び検索を行うことができる。またはこの特徴概念の表示を行うことができる。表示部510の図7に示す表示領域(2)では、カテゴリ[ソフトウェア]に含まれる概念のリストを表示してある。1つのキーワードに対して2つのグラフがあるが、上は頻度、下は相対頻度を表している。さらに頻度・相対頻度のどちらかで並び替えることができる。これらの表示部510における特徴的な概念の抽出及び表示の流れを以下により詳細に説明する。

【0048】

図8に表示領域(1)における、特徴的な概念の抽出及び表示のフローチャートを示す。ステップ820でカテゴリAとカテゴリBを選択する。ここでカテゴリA, Bはそれぞれ表示領域(1)でのx, y軸になる。後の計算で、カテゴリAの要素Axに関して特徴的なものを表示するので、比較したいカテゴリをBにセットする。ステップ840でカテゴリAを装置540に入力し、カテゴリA中に含まれる概念が頻度順に取り出される。またカテゴリBを装置540に入力し、カテゴリB中に含まれる概念が頻度順に取り出される。次にステップ850で、AとBの組合せ(A & B)が終了したかどうか判断される。終了していなければステップ855で、(Ai & Bj)を装置540に入力し、結果をPijに保存する。次にステップ860で、正規化が終了したかどうかを判断する。正規化が終了していなければステップ865において、装置530でBniを1として正規化する。すなわち($P'_{ij} = P_{ij} / B_{ni}$)となる。図6における表示例中の%表示がそれにあたる。ただし、1つの文章に複数の種類の概念が表れるので、これらの値を全部足しても1にはならない。次にステップ870で、相対頻度の計算が終了したかどうかを判断する。計算が終了していなければステップ875で、 P'_{ij} をy軸方向に比較して相対頻度を装置530で調べる。好適には P'_{ij} ($0 < i < a$)の分布の平均を調べ、各 P'_{ij} が平均の何倍になっているかを相対頻度とする。次に処理はステップ880へ進み、相対頻度の値によって、表示領域(1)の表示色を変え、2次元に表示する。次にステップ890で、目立っているカテゴリA中の概念faとカテゴリB中の概念fbの交点をクリックする。最後にステップ898で装置540により、カテゴリAの概念faとカテゴリBの概念fbでデータ集合が絞り込

まれる。

【0049】

図9に表示領域(2)における、本発明の別の特徴的な概念の抽出及び表示のフローチャートを示す。まずステップ915で表示領域(2)の左側でカテゴリAを選択する。次にステップ920で装置550により、カテゴリA中に含まれる概念が頻度順に取り出される。そしてステップ930で正規化が終了したかどうか判断される。終了していなければステップ940で相対頻度を計算する。ここでの相対頻度の計算は、ある概念が全体または任意の部分集合Xの中に含まれる割合 S_x と、任意の部分集合Yに含まれる割合 S_y の比(S_y/S_x)で求める。この値が大きいほど、部分集合Yに特徴的な概念と考える。ステップ930において正規化が終了した場合ステップ950で、表示領域(2)の右側に頻度と相対頻度を同時に表示し、頻度または相対頻度で並び替えて表示する。次にステップ960で表示領域(2)の右側で概念faを選択する。そして最後にステップ970で入力装置570を用いて選択(クリック)することで、装置540による絞込みが行える。

【0050】

上記2つの検索・抽出方法を組み合わせることにより、特徴的な概念を効果的に見つけることが可能となる。例えば表示領域(2)において、カテゴリ「月」について検索し「11月」を選択する(図示せず)。次に図10に示すように、表示領域(1)において「製品名(コンピュータの機種名)」を縦軸(比較したい対象)、「問題」を横軸にとる。すると「遅い」という「問題」に関して、特徴的な「製品名」が2つマークされている。そして、より相対頻度の高い「製品A」に注目し、「製品A」と「遅い」の交わっているところを選択(クリック)する。図11に示すように、絞り込まれた状態で表示領域(2)において、カテゴリ「ハードウェア」について見てみる。すると、頻度の上位2番目には「ハードディスク」があり、相対頻度も高く(7.18倍)、この製品Aに関しては「ハードディスク」が特有の問題を持っていることを推測できる。なお上位1番目は、本製品の製品番号などが該当するのでこれらは容易に無視できる。

【0051】

図 1 3 に上記操作手順を 1 つの画面内で行う本発明で最も特徴的な GUI の例を示す。図 1 3 において表示領域 (2) の左側でカテゴリ [月] を選ぶ (図示せず)。すると装置 5 5 0 により、カテゴリ [月] 中に含まれる概念が頻度順に取り出される。そして装置 5 3 0 により得られた概念の相対頻度が計算され、表示領域 (2) の右側に [月] に含まれる概念が表示される。次に、表示領域 (2) の右側で概念「11月」を選択する (図示せず)。これにより装置 5 4 0 により、カテゴリ [月] の「11月」でデータ集合が絞り込まれる。

【 0 0 5 2 】

次に表示領域 (1) でカテゴリ [問題] を X 軸に設定し、カテゴリ [機種名] を Y 軸に設定する。すると装置 5 5 0 でカテゴリ [問題] 中に含まれる概念が頻度順に取り出される。また装置 5 5 0 でカテゴリ [機種名] 中に含まれる概念が頻度順に取り出される。これらは検索・正規化され 2 次元に表示される。そして目立っているカテゴリ [問題] の「遅い」とカテゴリ [機種名] の「製品 A」の交点をクリックする。すると装置 5 4 0 により、カテゴリ [問題] の「遅い」とカテゴリ [機種名] の「製品 A」でデータ集合が絞り込まれる。

【 0 0 5 3 】

次に表示領域 (2) の左側でカテゴリ [ハードウェア] を選ぶ。装置 5 5 0 により、カテゴリ [ハードウェア] 中に含まれる概念が頻度順に取り出される。そして装置 5 3 0 により得られた概念の相対頻度が計算される。表示領域 (2) の右側に [ハードウェア] に含まれる概念が表示される。結局、上位 2 番目には「ハードディスク」があり相対頻度も高く、この製品に関しては「ハードディスク」が特有の問題であることが判明する。このように表示領域 (1) 及び表示領域 (2) を 1 画面内で組み合わせて操作することにより、特徴的な概念が容易に獲得でき、特定の製品に潜む根本的な問題などを簡単に見つけ出すことができる。

【 0 0 5 4 】

図 1 4 に本発明において使用されるデータ分析システムのハードウェア構成例を示す。システム 1 0 0 は、中央処理装置 (CPU) 1 とメモリ 4 とを含んでいる。CPU 1 とメモリ 4 は、バス 2 を介して、補助記憶装置としてのハードディスク装置 1 3 (または CD-ROM 2 6、DVD 3 2 等の記憶媒体駆動装置) と I

DEコントローラ25を介して接続してある。同様にCPU1とメモリ4は、バス2を介して、補助記憶装置としてのハードディスク装置30（またはMO28、CD-ROM29、DVD31等の記憶媒体駆動装置）とSCSIコントローラ27を介して接続してある。フロッピーディスク装置20はフロッピーディスクコントローラ19を介してバス2へ接続されている。好適にはラベル付きデータ160、辞書340、ラベル・概念付きデータ370はこれら補助記憶装置に記憶される。

【0055】

フロッピーディスク装置20には、フロッピーディスクが挿入され、このフロッピーディスク等やハードディスク装置13（またはCD-ROM26、DVD32等の記憶媒体）、ROM14には、オペレーティングシステムと協働してCPU等に命令を与え、本発明を実施するためのコンピュータプログラム、オペレーティングシステムのコード若しくはデータを記録することができ、メモリ4にロードされることによって実行される。これらコンピュータ・プログラムのコードは圧縮し、または、複数に分割して、複数の記録媒体に記録することもできる。

【0056】

システム100は更に、ユーザ・インターフェース・ハードウェアを備え、入力をするためのポインティング・デバイス（マウス、ジョイスティック等）7またはキーボード6や、ディスプレイ12を有することができる。好適にはポインティング・デバイス7を用いて、ディスプレイ12に表示された表示領域（1）、表示領域（2）の項目の選択、変更や、パラメータの入力をGUIで行う。また、パラレルポート16を介してプリンタを接続することや、シリアルポート15を介してモデムを接続することが可能である。このシステム100は、シリアルポート15およびモデムまたは通信アダプタ18（イーサネットやトークンリング・カード）等を介してネットワークに接続し、他のコンピュータ、サーバ等と通信を行う。本発明にデータ分析システムは、必要なデータベースを、通信回線を介して接続された外部のサーバ、WAN、LAN内にあるローカルなサーバなどに記憶してもよい。どちらにしても本発明の実施の制限となるものではない。またシ

リアルポート 1 5 若しくはパラレルポート 1 6 に、遠隔送受信機器を接続して、赤外線若しくは電波によりデータの送受信を行ってもよい。

【 0 0 5 7 】

スピーカ 2 3 は、オーディオ・コントローラ 2 1 によって D/A（デジタル／アナログ変換）変換されたサウンド、音声信号を、アンプ 2 2 を介して受領し、サウンド、音声として出力する。また、オーディオ・コントローラ 2 1 は、マイクロフォン 2 4 から受領した音声情報を A/D（アナログ／デジタル）変換し、システム外部の音声情報をシステムにとり込むことを可能にしている。ViaVoice（IBM商標）などのアプリケーションを用いて、本発明の GUI コマンド部の操作を音声コマンドによる操作で代用してもよい。さらにホームページ・リーダー（IBM商標）などアプリケーションを用いて、表示された特徴的概念を有する検索結果などを音声で読み上げるようにしてもよい。

【 0 0 5 8 】

このように、本発明のデータ分析システムは、通常のパーソナルコンピュータ（PC）やワークステーション、ノートブック PC、パームトップ PC、ネットワークコンピュータ、コンピュータを内蔵したテレビ等の各種家電製品、通信機能を有するゲーム機、電話、FAX、携帯電話、PHS、電子手帳、等を含む通信機能有する通信端末、または、これらの組合せによって実施可能であることを容易に理解できるであろう。ただし、これらの構成要素は例示であり、その全ての構成要素が本発明の必須の構成要素となるわけではない。

【 0 0 5 9 】

【発明の効果】

本発明により、大量データから有効な知識を獲得する方法およびシステムが提供される。また有効な知識を獲得するための、ユーザビリティに優れた GUI による分析方法およびシステムが提供される。

【図面の簡単な説明】

【図 1】

本発明のデータ分析システム概要を示すブロック図である

【図 2】

データ変換部のフローチャートである。

【図 3】

概念抽出部のブロック図である。

【図 4】

概念抽出部のフローチャートである。

【図 5】

検索・特徴抽出部のブロック図である。

【図 6】

表示部における表示領域（１）の例である。

【図 7】

表示部における表示領域（２）の例である。

【図 8】

表示領域（１）における概念の抽出及び表示のフローチャートである。

【図 9】

表示領域（２）における概念の抽出及び表示のフローチャートである。

【図 1 0】

表示領域（１）の典型的な表示例である。

【図 1 1】

表示領域（２）の典型的な表示例である。

【図 1 2】

概念抽出部の処理を具体的な文を用いて説明した図である

【図 1 3】

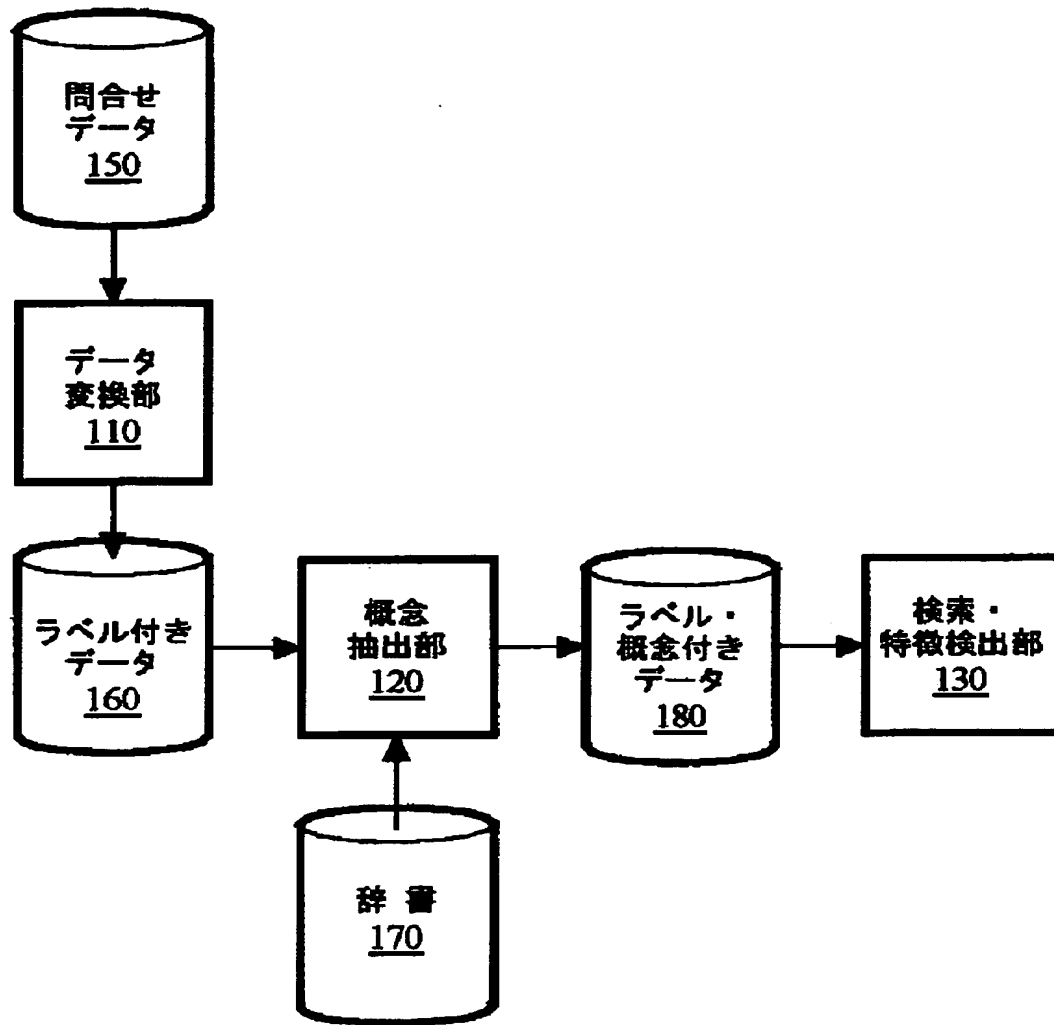
表示領域（１）と表示領域（２）を含む G U I 画面の例である。

【図 1 4】

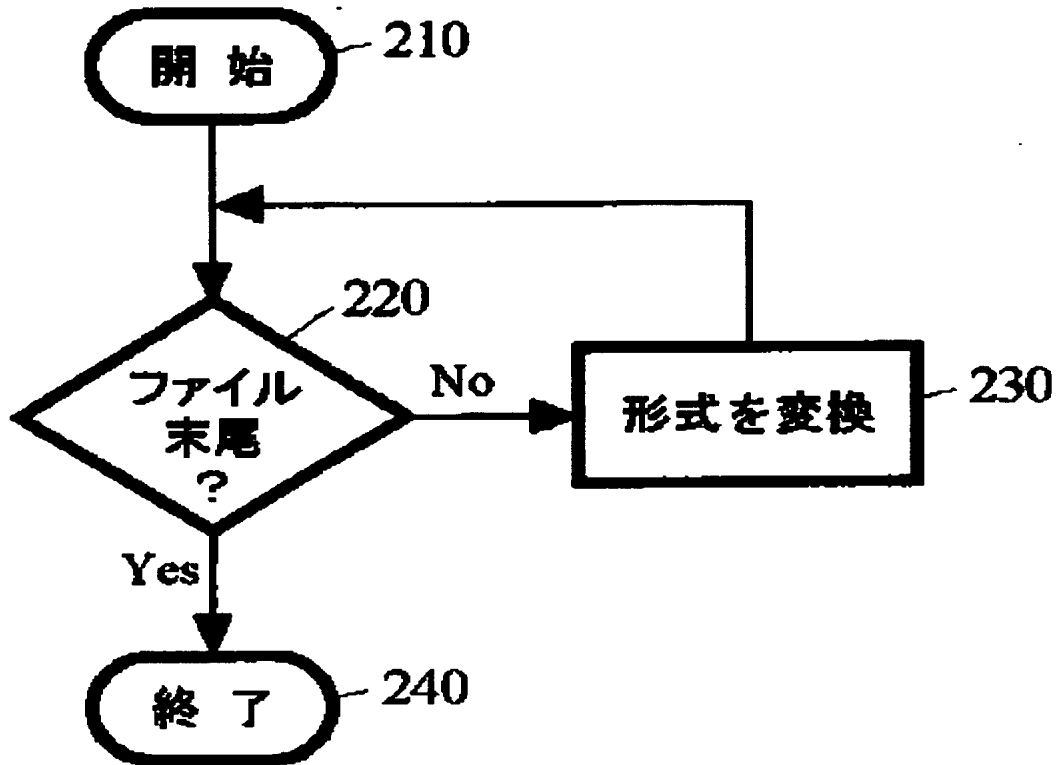
本発明に用いるハードウェアの実施例である。

【書類名】 図面

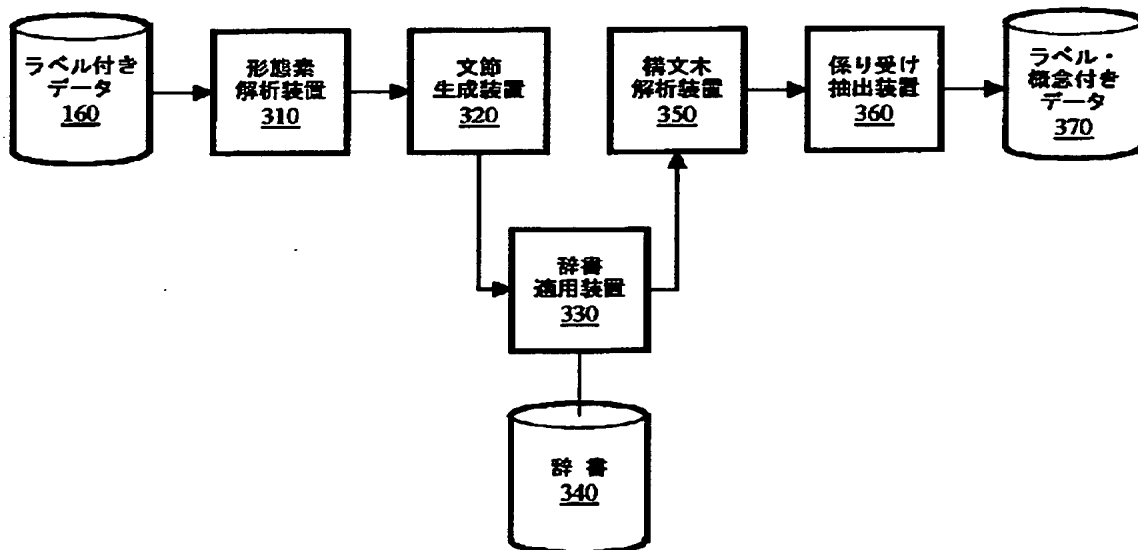
【図 1】



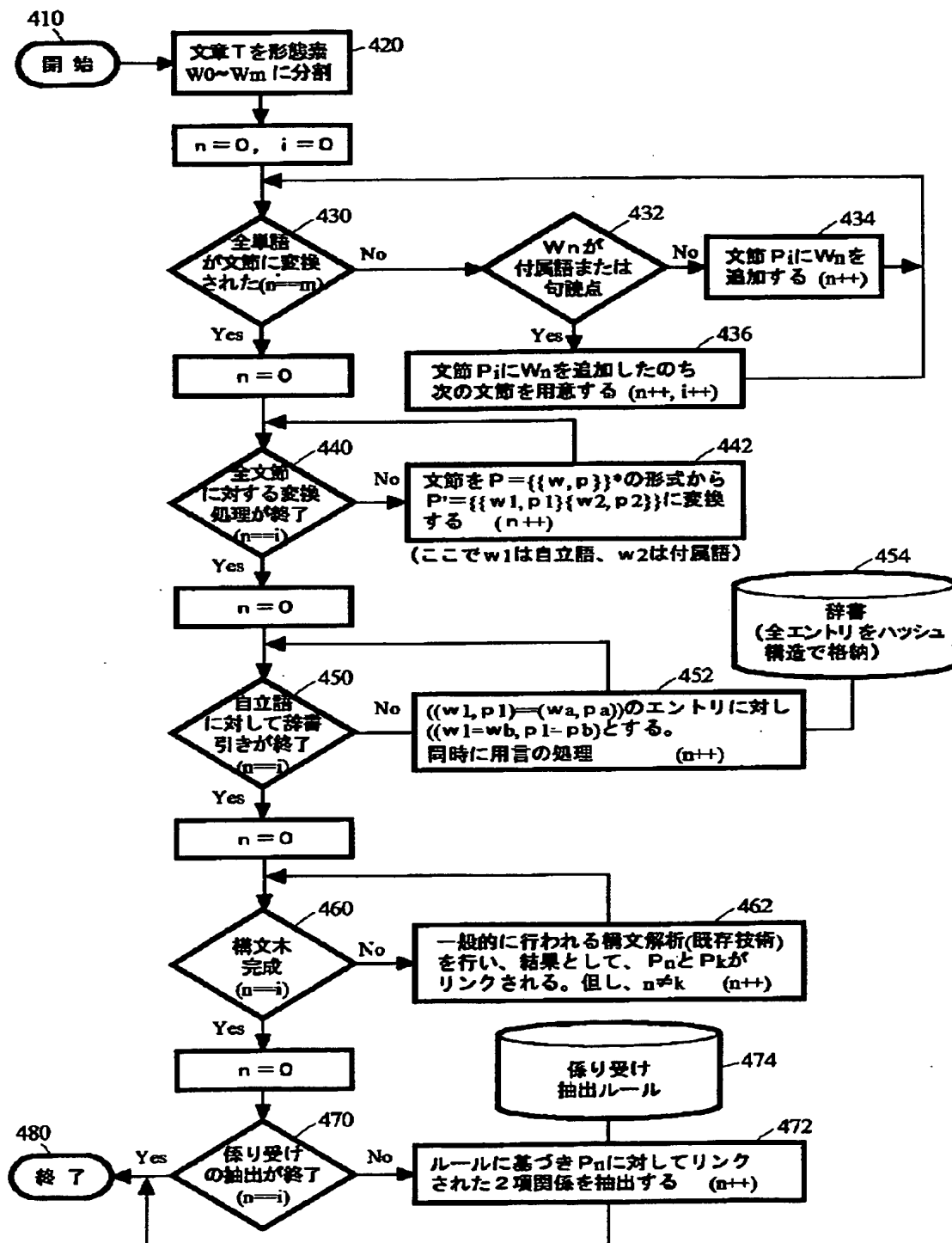
【図 2】



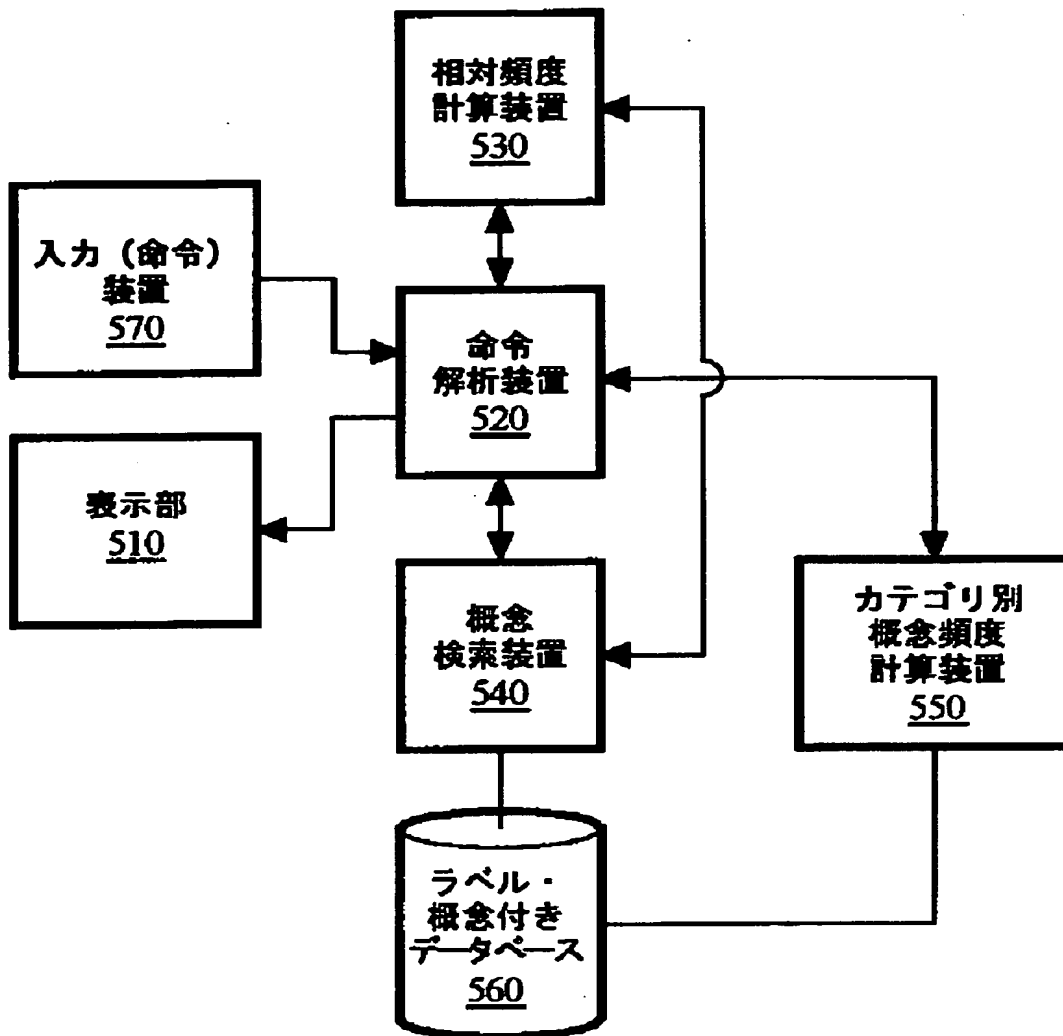
【図 3】



【図 4】



【図 5】



【図 6】

fa
↓ カテゴリ A

		A1	A2	A3	A4	A5	A6	A7
fb →	B1	1720 (89.6%)	56 (2.9%)	118 (6.1%)	7 (0.3%)	1 (0.0%)	7 (0.3%)	1 (0.0%)
	B2	1255 (90.5%)	21 (1.5%)	100 (7.2%)	4 (0.2%)	2 (0.1%)	2 (0.1%)	2 (0.1%)
	B3	1072 (95.0%)	28 (2.4%)	18 (1.5%)	5 (0.4%)	0 (0.0%)	3 (0.2%)	1 (0.0%)
カテゴリ B	B4	501 (90.1%)	9 (1.6%)	40 (7.1%)	1 (0.1%)	3 (0.5%)	2 (0.3%)	0 (0.0%)
	B5	649 (93.2%)	26 (3.7%)	17 (2.4%)	1 (0.1%)	0 (0.0%)	3 (0.4%)	0 (0.0%)
	B6	682 (95.5%)	13 (1.8%)	15 (2.1%)	3 (0.4%)	0 (0.0%)	1 (0.1%)	0 (0.0%)
	B7	676 (96.9%)	11 (1.5%)	5 (0.7%)	3 (0.4%)	1 (0.1%)	1 (0.1%)	0 (0.0%)
	B8	673 (90.7%)	34 (4.5%)	26 (3.5%)	4 (0.5%)	0 (0.0%)	5 (0.6%)	0 (0.0%)
	B9	528 (74.3%)	134 (18.8%)	10 (1.4%)	14 (1.9%)	2 (0.2%)	22 (3.0%)	0 (0.0%)
	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

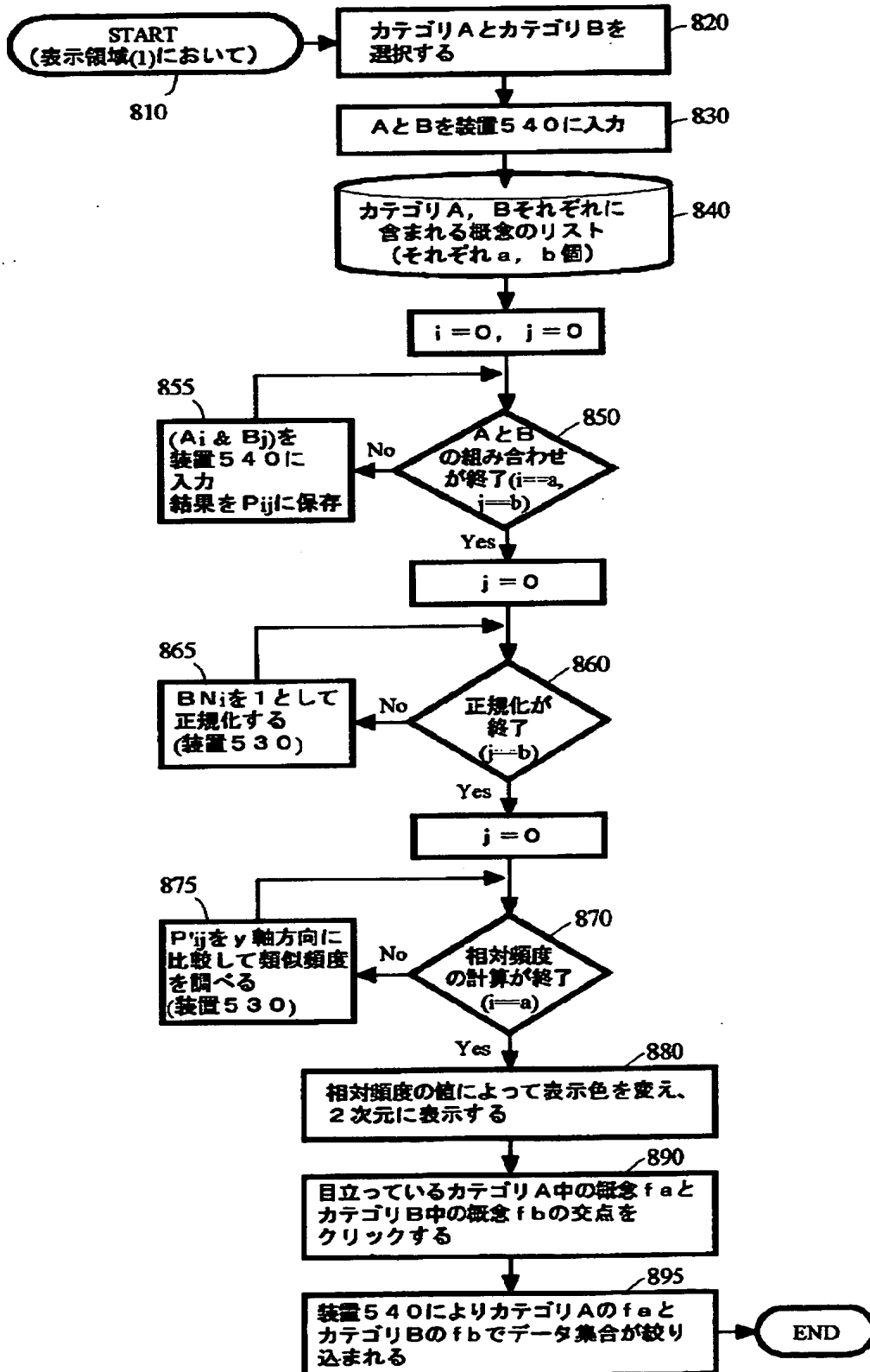
表示領域 (1)

【図 7】

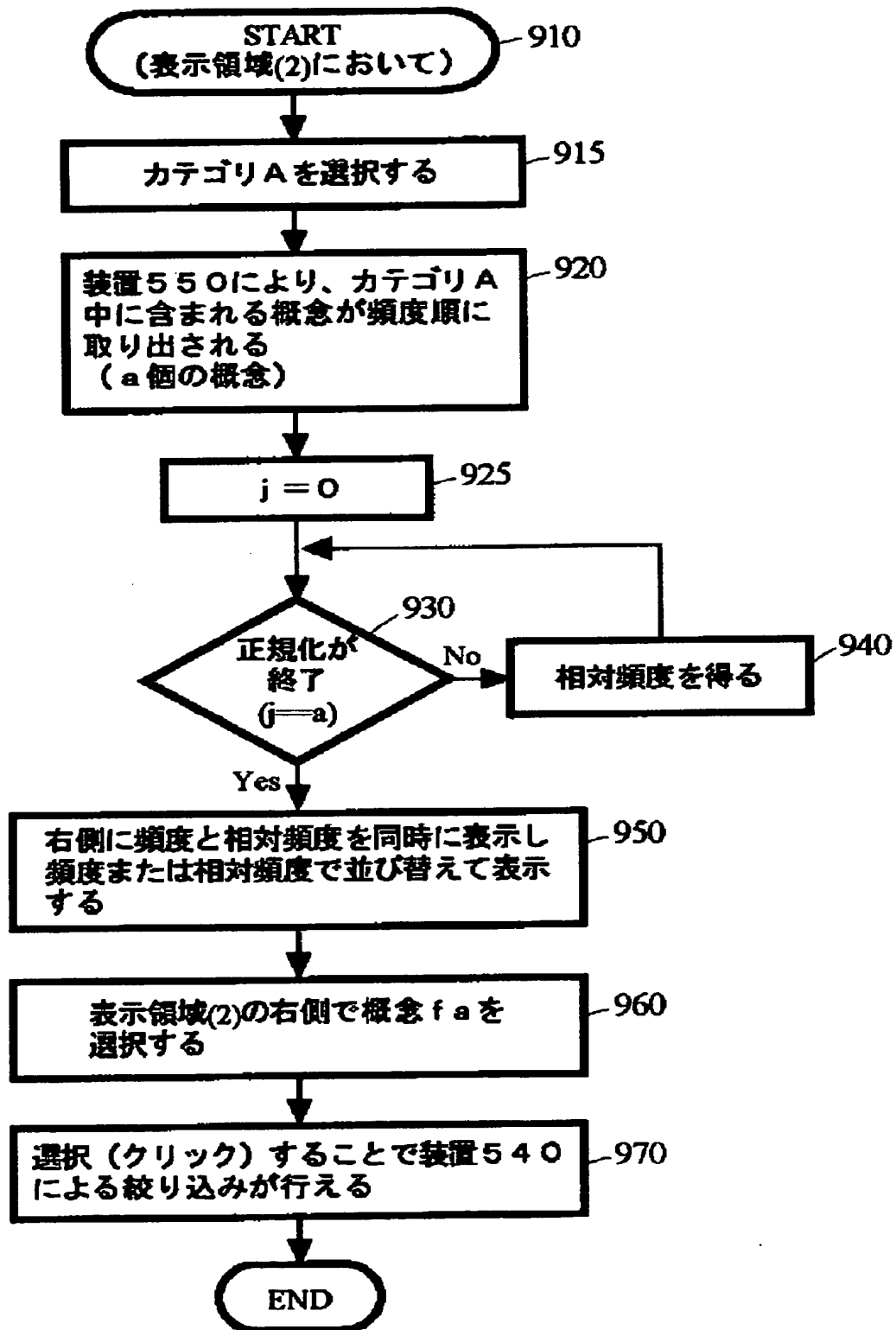
	件数・相対頻度	ソート ○ 件数 ◎ 相対	スナップショット
カテゴリ B	ソフトウェア	1255	WINDOWS98
	ハードウェア	120	WINDOWS98
	専門用語	118	アップグレード
	コマンド	77	リカバリー
	対象コンボ	80	メッセージ
	問題種別	43	WINDOWS
	CALL種別	34	ドライバ
	回答・対応種別	27	ソフト
	機種名	27	起動時
	解決時間 (分)	27	アプリケーション
	CALL回数	19	
	対応担当者数		
	対応チーム数		
	名詞		
	固有名詞		
	その他		
	人名		
	組織名		

表示領域 (2)

【図 8】



【図9】



【図 1 0】

						遅い	
	19 (2.58%)	6 (0.82%)	9 (1.22%)	6 (0.82%)	2 (0.27%)	2 (0.27%)	3 (0.41%)
	18 (2.02%)	7 (0.79%)	18 (2.02%)	4 (0.45%)	4 (0.45%)	3 (0.34%)	3 (0.34%)
	21 (2.75%)	14 (1.83%)	15 (1.97%)	12 (1.57%)	12 (1.57%)	14 (1.83%)	7 (0.92%)
	57 (5.76%)	12 (1.21%)	11 (1.11%)	8 (0.81%)	3 (0.31%)	1 (0.10%)	5 (0.51%)
	55 (4.8%)	20 (1.75%)	7 (0.61%)	11 (0.96%)	2 (0.17%)	1 (0.95%)	5 (0.44%)
製品 A	20 (2.58%)	15 (1.94%)	14 (1.81%)	8 (1.03%)	6 (0.78%)	20 (2.58%)	6 (0.78%)
	15 (2.11%)	18 (2.53%)	12 (1.69%)	9 (1.27%)	1 (0.14%)	5 (0.70%)	3 (0.42%)
	17 (2.29%)	10 (1.35%)	33 (4.45%)	6 (0.81%)	1 (0.13%)	2 (0.27%)	2 (0.27%)
	15 (2.76%)	7 (1.29%)	8 (1.47%)	5 (0.92%)	2 (0.37%)	0 (0.00%)	3 (0.55%)
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

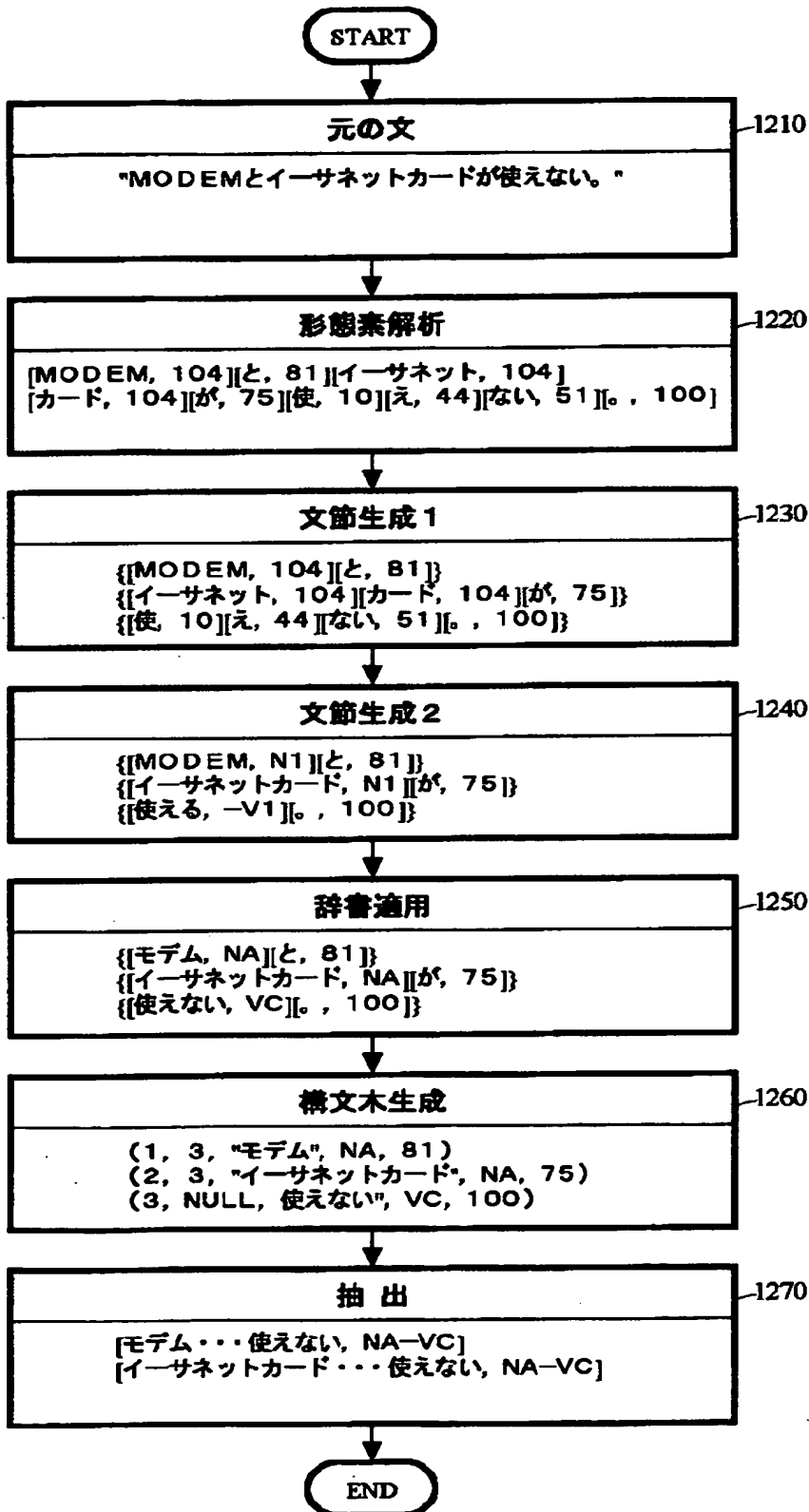
表示領域 (1)

【図 1 1】

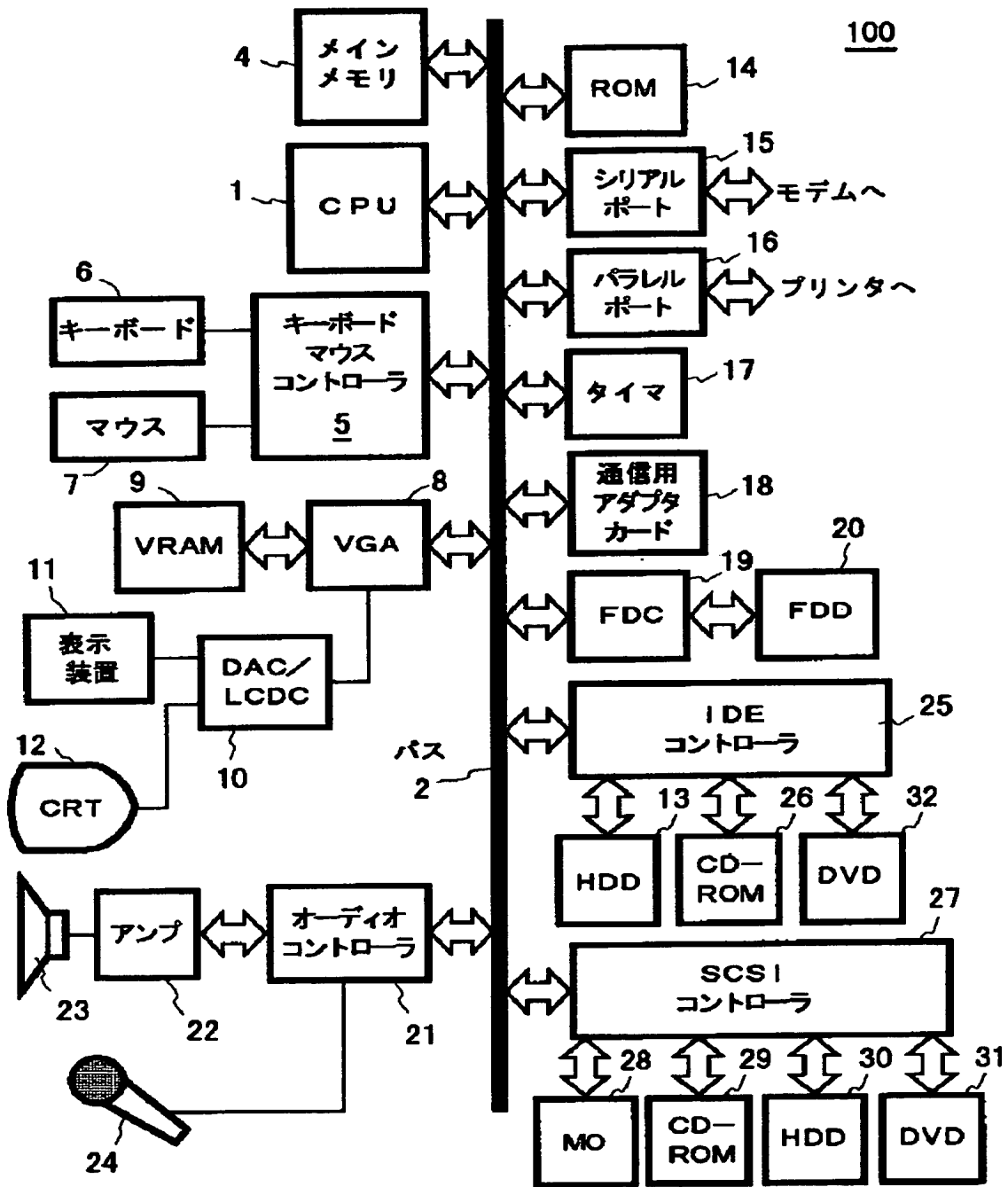
件数・相対頻度	ソート	◎ 件数	○ 相対	○ 項目	スナップショット
ソフトウェア	187	187	100%	-----	
ハードウェア	11	11	5.9%	ハードディスク	
専門用語	7.18	7.18	3.8%	-----	
コマンド	3.1	3.1	1.6%	-----	
対象コンポ	2.0	2.0	1.0%	-----	
問題種別	2.0	2.0	1.0%	-----	
CALL 種別	2.0	2.0	1.0%	-----	
回答・対応種別	1.0	1.0	0.5%	-----	
機種名	1.0	1.0	0.5%	-----	
解決時間 (分)	1.0	1.0	0.5%	-----	
CALL 回数	1.0	1.0	0.5%	-----	
対応担当者数	1.0	1.0	0.5%	-----	
対応チーム数	1.0	1.0	0.5%	-----	
名詞	1.0	1.0	0.5%	-----	
固有名詞	1.0	1.0	0.5%	-----	
その他				-----	
人名				-----	
組織名				-----	

表示領域 (2)

【図 12】



【図 1 4】



【書類名】 要約書

【要約】

【課題】 大量データから有効な知識を獲得する方法およびシステムを提供することである。

【解決手段】

本発明は文書を含む大量のデータから特異な特徴を有する概念を抽出することにより、有効な知識を獲得する方法およびシステムを提供する。本発明は概念抽出装置と特徴的概念抽出装置とを有する。前記概念抽出は文書データを含むデータからカテゴリ別の概念を抽出する。前記特徴的概念抽出装置は前記抽出した概念の中から特徴的な概念を抽出する装置であって、前記カテゴリ別の概念において、同一カテゴリに属する概念のうち、対応する別のカテゴリに属する概念の中で占める割合が既定値を超えている概念を抽出する。

【選択図】 図 1

認定・付加情報

特許出願の番号	平成11年 特許願 第241057号
受付番号	59900830001
書類名	特許願
担当官	第七担当上席 0096
作成日	平成11年 9月 2日

<認定情報・付加情報>

【提出日】	平成11年 8月27日
【特許出願人】	
【識別番号】	390009531
【住所又は居所】	アメリカ合衆国10504、ニューヨーク州 アーモンク (番地なし)
【氏名又は名称】	インターナショナル・ビジネス・マシーンズ・コーポレーション
【代理人】	申請人
【識別番号】	100086243
【住所又は居所】	神奈川県大和市下鶴間1623番地14 日本アイ・ビー・エム株式会社 大和事業所内
【氏名又は名称】	坂口 博
【選任した代理人】	
【識別番号】	100091568
【住所又は居所】	神奈川県大和市下鶴間1623番地14 日本アイ・ビー・エム株式会社 大和事業所内
【氏名又は名称】	市位 嘉宏

出 願 人 履 歴 情 報

識別番号 [390009531]

1. 変更年月日 1990年10月24日

[変更理由] 新規登録

住 所 アメリカ合衆国10504、ニューヨーク州 アーモンク (番地なし)

氏 名 インターナショナル・ビジネス・マシーンズ・コーポレイション